DOCUMENT RESUME

ED 193 302 TM 800 617

AUTHOR Mayekawa, Shin-ichi; Haebara, Tomokazu
TITLE Estimation of the Reliability of a Test Consistency of More Than Three Congeneric Parts. Iowa Testing Programs Occasional Papers, Number 28.
REPORT NO ITPOP28
PUB DATE May 80
NOTE 17p.; Paper presented at the Psychometric Society Meetings (Iowa City, IA, May, 1980).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests; *Least Squares Statistics; Mathematical Models; *Test Reliability
IDENTIFIERS *Estimation (Mathematics); Reliability Formulas

ABSTRACT

A least squares approach to estimating the reliability of a measure consisting of more than three content homogeneous or congeneric parts is proposed. The advantages of this method over a more indirect approach in which certain parts of a measure are combined to use Kristof's or Feldt's coefficients are examined. One hundred four-part tests were created by randomly assigning certain items from a 48-item mathematics achievement test to each of the four parts. Data on 400 examinees who took the test were divided into five subsets. Seven estimates of reliability were computed for 100 random splits for each subset of data. Some of the main results included: (1) there is a consistent order in the magnitudes of the means of the coefficients, regardless of sample size; (2) values of the average of the Kristof coefficients are greater than unity when sample size is small; (3) the proposed coefficient has smaller standard deviation and range than the Kristof coefficient; and (4) the Feldt coefficient tends to have smaller standard deviation and range than the proposed coefficient, and depends on the assumption that the error variance of a part is proportional to its length. (Author/RL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED193302

U S DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

IOWA TESTING PROGRAMS OCCASIONAL PAPERS

Number 28 - August 1980

Estimation of the Reliability of a Test Consisting of More Than Three Congeneric Parts

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
T. Haebara
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Shin-ichi Mayekawa

and

Tomokazu Haebara

The University of Iowa

TM 800617

Paper presented at the
Psychometric Society Meetings
Iowa City, May 1980

Estimation of the Reliability of a Test Consisting of More Than Three Congeneric Parts

The purpose of this study is to propose a least squares approach to estimating the reliability of a measure consisting of more than three content homogeneous or congeneric parts. This approach is a natural extension of Kristof's and Feldt's reliability coefficients that deal with the three-part case and the two-part case, respectively. Several advantages of the method over a more indirect approach in which certain parts of a measure are combined to use Kristof's or Feldt's coefficients are claimed based on empirical results.

Introduction

Estimation of the reliability of a measure by a single administration has long been investigated. Among those techniques developed for this purpose, Cronbach's (1951) coefficient alpha is probably the most widely used reliability coefficient. However, Cronbach's alpha and Guttman's λ_2 (1945), which is sometimes preferred as a better lower bound for the reliability, are appropriate coefficients only when the assumption of essential tau equivalence is at least approximately satisfied for the constituent parts of the measure.

If, for example, a measure of interest is defined as the sum of the ratings given by several judges, or if the parts of a test are not equal in length, essential tau equivalence is not a reasonable assumption. Lord and Novick (1968, Section 9.12) provided a general orientation toward estimating the reliabilities of such measures. Under this orientation, Kristof (1974)

and Feldt (1975) developed formulas for the reliabilities in the cases of three-part tests and two-part tests, respectively. Both Kristof and Feldt adopted the assumption of perfectly correlated true part-test scores, which is referred to as the assumption of content homogeneous or congeneric parts.

When a measure consists of more than three congeneric parts, one could reduce the number of parts by combining certain parts to use Kristof's or Feldt's coefficient. Although an indeterminacy problem may result from ambiguity about which parts to combine, this problem may be overcome by taking the average of coefficients obtained from all possible ways of combining the parts. In this study, however, we propose a more direct approach to estimating the reliability of such a measure. This direct method will be evaluated later in this paper by comparing the empirical results obtained from it and those obtained by taking the average of Kristof's coefficients and the average of Feldt's coefficients.

Derivation

Let the observed score be represented by X, which is the sum of n part scores X_j (j=1,2,...,n), and the corresponding true scores and error scores by T, $T_j + C_j$ (j=1,2,...,n), E, and E_j (j=1,2,...,n). And let the relative functional length of each part be represented by λ_j (j=1,2,...,n). Assuming content homogeneity of the parts, we can write the basic relationships as follows:

$$X = \sum_{j=1}^{n} X_j = T + E,$$

$$E = \sum_{j=1}^{n} E_j$$

$$X_j = T_j + C_j + E_j,$$

$$T_j = \lambda_j T,$$

$$\sum_{j=1}^{n} \lambda_j = 1, \qquad (1)$$

and

$$\sum_{j=1}^{n} C_j = 0.$$

Then the observed score variances and covariances of the n parts are

$$\sigma_j^2 = \lambda_j^2 \sigma_T^2 + \sigma_{Ej}^2, \quad j=1,2,\ldots,n;$$

and

$$\sigma_{ij} = \lambda_i \lambda_j \sigma_T^2, \quad i, j=1,2,\ldots,n, \quad i \neq j.$$

These relationships and Equation (1) constitute a system of $(n+n(n-1)/2+1)$ equations for $(2n+1)$ unknowns (σ_T^2, λ_j, and σ_{Ej}^2, $j=1,2,\ldots,n$).

Thus it is seen that if the number of parts is greater than three, we have an overdetermined system of equations which does not have a unique solution. To avoid this overdeterminacy, we apply the method of Least Squares.

Let us define the Least Squares (L.S.) criterion as follows:

$$Q = \sum_{j=1}^{n} \{\sigma_j^2 - (\lambda_j^2 \sigma_T^2 + \sigma_{Ej}^2)\}^2 + \sum_{i \neq j} \{\sigma_{ij} - \lambda_i \lambda_j \sigma_T^2\}^2. \qquad (2)$$

Now our aim is to find the values σ_T^2, σ_{Ej}^2's, and λ_j's that minimize this criterion. Rewriting (2) in a matrix form, we have

$$Q = \mathrm{tr}\{(\Sigma - (\sigma_T^2 \underline{\lambda}\underline{\lambda}' + D))^2\}$$

$$= \mathrm{tr}\{(\Sigma - (\underline{a}\underline{a}' + D))^2\}, \tag{3}$$

where Σ is the observed variance-covariance matrix, and

$$\underline{\lambda} \equiv (\lambda_1, \lambda_2, \ldots, \lambda_n)',$$

$$D \equiv \begin{bmatrix} \sigma_{E1}^2 & & & \\ & \sigma_{E2}^2 & & \\ & & \ddots & \\ & & & \sigma_{En}^2 \end{bmatrix},$$

and

$$\underline{a} \equiv \sigma_T \underline{\lambda} \,. \tag{4}$$

The partial derivatives of the L.S. criterion with respect to $\underline{a}$ and D are given by

$$\frac{\partial Q}{\partial \underline{a}} = -4\{(\Sigma - D)\underline{a} - \underline{a}\underline{a}'\underline{a}\} \tag{5}$$

and

$$\frac{\partial Q}{\partial D} = -2\{\mathrm{diag}(\Sigma - aa') - D\}. \tag{6}$$

Setting these equations equal to zero, we get

$$(\Sigma - D)\underline{a} = \underline{a}\underline{a}'\underline{a}, \tag{7}$$

and

$$D = \mathrm{diag}(\Sigma - \underline{a}\underline{a}'). \tag{8}$$

When we rewrite (7) as

$$(\Sigma - D)\underline{a} = v\underline{a}, \tag{9}$$

where

$$v = \underline{a}'\underline{a}, \tag{10}$$

it becomes clear that $\underline{a}$ is an eigen vector of the matrix $(\Sigma - D)$, which is normalized so that its squared length is equal to the corresponding eigen value v.

Using Equation (9), we can rewrite the L.S. criterion (3) as

$$\begin{aligned} Q &= \mathrm{tr}\{(\Sigma - D)^2\} - 2\mathrm{tr}\{(\Sigma - D)\underline{a}\underline{a}'\} + v\ \mathrm{tr}\{\underline{a}\underline{a}'\} \\ &= \mathrm{tr}\{(\Sigma - D)^2\} - 2v\ \mathrm{tr}\{\underline{a}\underline{a}'\} + v\ \mathrm{tr}\{\underline{a}\underline{a}'\} \\ &= \mathrm{tr}\{(\Sigma - D)^2\} - v^2, \end{aligned} \tag{11}$$

which shows that v should be the largest eigen value of $(\Sigma - D)$ so that the minimum of Q may be obtained.

Denoting the eigen vector of unit length corresponding to v by $\underline{p}$, the solution of Equation (9) is given by

$$\hat{\underline{a}} = \sqrt{v}\ \underline{p}. \tag{12}$$

And, by Equation (1),

$$\hat{\underline{\lambda}} = \frac{1}{\sum_{j=1}^{n} p_j}\ \underline{p}$$

and

$$\hat{\sigma}_T^2 = v\left(\sum_{j=1}^{n} p_j\right)^2. \tag{13}$$

where $\hat{}$ denotes the L.S. estimator. The solution of Equation (8) may now be be written as

$$\hat{D} = \mathrm{diag}(\Sigma - \hat{\underline{a}}\hat{\underline{a}}'). \tag{14}$$

In the actual estimation procedure, the eigen vector of $(\Sigma - D)$ must be

found by successively updating the elements of D by (14) until we arrive at the stationary point where

$$\{\text{off}(\Sigma) + \text{diag}(\hat{\underline{a}}\hat{\underline{a}}')\}\hat{\underline{a}} = \hat{\underline{a}}\hat{\underline{a}}'\hat{\underline{a}} \tag{15}$$

is satisfied. This type of successive procedure is called an Alternating Least Squares method (Young, de Leeuw, and Takane, 1976). The above derivation is essentially the same as that for the minimum residual method (Comrey, 1962) in factor analysis. Actually, that the present problem is a specialized problem in factor analysis was previously pointed out by Lord and Novick (1968). Finally, the reliability coefficient ρ is estimated by

$$\hat{\rho} = \frac{\hat{\sigma}_T^2}{\sigma_X^2}, \tag{16}$$

where $\hat{\sigma}_T^2$ is given by (13).

If the measure consists of three parts, this L.S. solution is equivalent to Kristof's solution; that is, Kristof's solution gives the absolute minimum of Q in (3). However, if the measure consists of only two parts, the solution is not unique; that is, Equation (9) is reduced to a single equation,

$$\sigma_{12} = \sigma_T^2 \lambda_1 \lambda_2$$

or

$$\sigma_{12} = a_1 a_2,$$

which cannot be solved for the unknowns without an additional assumption such as that employed by Feldt (1975).

Evaluation

To investigate the characteristics of the proposed reliability coefficient (16), we created one hundred four-part tests by randomly assigning a certain number of items from a 48-item mathematics achievement test to each of the four parts. The ratios of nominal length (i.e., number of items) of parts were 1:1:1:1, 1:1:2:2, 3:5:7:9, and 1:3:5:7. Twenty-five random four-part tests were created for each split ratio. The four split ratios used here are intended to cover a plausible range of the variability in the functional lengths of four judges, four-part tests, etc. Since, in the present context, the total observed score of an individual is always his/her score on the same 48-item test, regardless of the way of dividing the test into parts, and since the reliability of that total score is sought for, it is desired that the 100 estimates of the reliability derived from the 100 randomly split tests be close to each other and to the true reliability coefficient.

Data on 400 examinees were available on the test. From this data set, we created five subsets of the data with sample sizes of 10, 25, 50, 100, and 400, in such a way that the examinees in samples of the smaller size are always included in samples of the larger size.

For each of 100 random splits for each of the five sample sizes, we computed the following seven estimates of the reliability:

r_{MH} : The proposed coefficient.

r_{K1} : The average of the Kristof coefficients obtained from all of the six possible ways of combining data to make three parts.

r_{K2} : The Kristof coefficient obtained by combining the two parts with the smallest nominal lengths.

r_{F1} : The average of the Feldt coefficients obtained from all of the seven possible ways of combining data to make two parts.

r_{F2} : The average of the Feldt coefficients obtained from all of the three possible ways of making two parts so that each combined part is the sum of two original parts.

r_{α} : Cronbach's coefficient alpha.

r_{G} : Guttman's λ_2.

The mean, standard deviation, maximum, minimum, and range of the 100 estimates of each coefficient were computed for each sample size and are shown in Table 1. Some of the main results are summarized as follows:

1. There is a consistent order in the magnitudes of the means of the coefficients, regardless of the sample size. r_{MH} always falls between Feldt's coefficients (r_{F1} and r_{F2}) and Kristof's coefficients (r_{K1} and r_{K2}), the latter being the largest.
2. Some values of r_{K1} are greater than unity when sample size is small (N = 10 and 25).
3. r_{MH} always has smaller SD and range than r_{K1} and r_{K2}, but r_{F1} and r_{F2} tend to have smaller SD and range than r_{MH}.

The first result given above implies an advantage of r_{MH} over the other coefficients concerning bias of estimation: Sedere and Feldt (1977)

Table 1

Summary Statistics for Seven Coefficients

	Mean	SD	Maximum	Minimum	Range
Sample Size = 10					
r_{MH}	.9206	.0262	.9686	.8518	.1168
r_{K1}	.9399	.0706	1.4901	.8434	.6467
r_{K2}	.9226	.0306	.9950	.8405	.1545
r_{F1}	.9144	.0241	.9640	.8422	.1218
r_{F2}	.9162	.0230	.9632	.8548	.1084
r_{α}	.8665	.0454	.9495	.7196	.2299
r_{G}	.8961	.0293	.9530	.8249	.1281
Sample Size = 25					
r_{MH}	.8864	.0241	.9482	.8231	.1251
r_{K1}	.8946	.0306	1.0483	.8295	.2188
r_{K2}	.8889	.0279	.9806	.8179	.1627
r_{F1}	.8825	.0213	.9432	.8255	.1177
r_{F2}	.8838	.0212	.9451	.8257	.1194
r_{α}	.8353	.0435	.9332	.7154	.2178
r_{G}	.8633	.0295	.9379	.7553	.1826
Sample Size = 50					
r_{MH}	.8934	.0153	.9313	.8425	.0888
r_{K1}	.8958	.0163	.9520	.8553	.0967
r_{K2}	.8947	.0174	.9557	.8448	.1109
r_{F1}	.8923	.0135	.9318	.8602	.0716
r_{F2}	.8926	.0135	.9312	.8618	.0694
r_{α}	.8487	.0384	.9136	.7392	.1744
r_{G}	.8734	.0227	.9156	.7853	.1303

Table 1 (Continued)

	Sample Size = 100				
	Mean	SD	Maximum	Minimum	Range
r_{MH}	.9172	.0095	.9349	.8893	.0456
r_{K1}	.9181	.0099	.9446	.8839	.0607
r_{K2}	.9175	.0106	.9356	.8883	.0473
r_{F1}	.9170	.0084	.9346	.8966	.0380
r_{F2}	.9171	.0083	.9346	.8959	.0387
r_{α}	.8735	.0356	.9337	.7906	.1431
r_{G}	.8980	.0182	.9342	.8470	.0872

	Sample Size = 400				
	Mean	SD	Maximum	Minimum	Range
r_{MH}	.8939	.0071	.9095	.8777	.0318
r_{K1}	.8941	.0075	.9090	.8791	.0299
r_{K2}	.8939	.0080	.9094	.8741	.0353
r_{F1}	.8931	.0068	.9103	.8785	.0318
r_{F2}	.8932	.0068	.9098	.8783	.0315
r_{α}	.8514	.0336	.9063	.7721	.1342
r_{G}	.8750	.0162	.9064	.8327	.0737

reported, based on a large-scale Monte Carlo study, that Kristof's coefficient is positively biased and Feldt's coefficient is negatively biased. Thus, although our result cannot directly confirm that r_{MH} is less biased than both Feldt's and Kristof's coefficients, the differences of r_{MH} with those coefficients are in directions toward reducing estimation bias.

The second and third results show that r_{K1} and r_{K2} are not as good reliability coefficients as r_{MH}, r_{F1} and r_{F2}. That the Kristof coefficient sometimes takes a value greater than unity is also reported in Sedere and Feldt (1977). Among r_{MH}, r_{F1}, and r_{F2}, r_{MH} tends to have a greater variability, but whether r_{F1} and r_{F2} are better coefficients than r_{MH} or not is an open question. As noted earlier, the Feldt coefficient depends on an additional assumption which is not necessary either with the Kristof coefficient or with r_{MH}. The assumption in question is, "The error variance of a part is proportional to its length." A reasonable conclusion would be that r_{MH} should be preferred to r_{F1} and r_{F2} as long as the validity of that assumption is in doubt.

In the present study, we did not analyze measures with more than four parts. However, the proposed method for estimating the reliability can also be used with such measures. And further empirical investigation with those measures and Monte Carlo studies similar to Sedere and Feldt (1977) will enhance our understanding on the characteristics of the proposed coefficient.

References

Comrey, A. L. The minimum residual method of factor analysis. Psychological Reports, 1962, 11, 15-18.

Cronbach, L. J. Coefficient alpha and the internal structure of tests. Psychometrika, 1951, 16, 297-334.

Feldt, L. S. Estimation of the reliability of a test divided into two parts of unequal length. Psychometrika, 1975, 40, 557-561.

Guttman, L. A basis for analyzing test retest reliability. Psychometrika, 1945, 10, 255-283.

Kristof, W. Estimation of reliability and true score variance from a split of a test into three arbitrary parts. Psychometrika, 1974, 39, 491-499.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Sedere, M. U., & Feldt, L. S. The sampling distributions of the Kristof reliability coefficient, the Feldt coefficient, and Guttman's lambda-2. Journal of Educational Measurement, 1977, 14, 53-62.

Young, F. W., de Leeuw, J., & Takane, Y. Multiple and canonical regression at a mix of quantitative and qualitative variables: An alternating least squares method with optimal scaling features. Psychometrika, 1976, 41, 504-529.

Acknowledgment

The problem discussed here was first suggested to us by Dr. Leonard S. Feldt. We are also indebted to Dr. William E. Coffman for providing the data used in this study, and to Ms. Kary Jacobs for typing. The main part of RELIABLE, a computer program developed for performing the computations discussed in the paper, uses a routine for the minimum residual method given in Inshi Bunseki Ho (Factor Analysis) by Dr. Sukeyori Shiba. Description and listing of the program are available from the first author.

IOWA TESTING PROGRAMS

OCCASIONAL PAPERS

1. Coffman, William E. Moratorium? What Kind? (ED 099 408)

2. Coffman, William E., & Mathews, Walter M. Narrative Reports and the Iowa Tests of Basic Skills.

3. Forsyth, Robert A., & Feldt, Leonard S. On the Validity of the I.T.E.D. as an Aid in Program Evaluation. (ED 142 558)

4. Gulliksen, Harold O. Applications of Psychological Scaling Methods.

5. Lee, Lai-Min Paul, & Coffman, William E. A Study of the "I Don't Know" Response in Multiple-Choice Tests. (ED 141 371)

6. Novick, Melvin R. A Course in Bayesian Statistics. (out of print) (In: American Statistician, May 1975, Vol. 29, No. 2, pp. 94-98)

7. Hood, Joyce, Leslie, Lauren, & Kendall, Janet Ross. Instructional Effectiveness in Primary Grade Reading: A Pilot Study.

8. Petersen, Nancy S., & Novick, Melvin R. An Evaluation of Some Models for Test Bias. (out of print) (In: Journal of Educational Measurement, Spring 1976, Vol. 13, No. 1, pp. 3-31)

9. Kendall, Janet Ross, & Hood, Joyce. Instructional Practices and Easy-to-Change Surrounding Conditions Variables in Effective Primary Reading Programs. (ED 123 557)

10. Petersen, Nancy S. An Expected Utility Model for "Optimal" Selection. (out of print) (In: Journal of Educational Statistics, Winter 1976, Vol. 1, No. 4, pp. 333-358)

11. Lindley, D. V. The Effect of Ethical Design Considerations on Statistical Analysis.

12. Lindley, D. V. A Class of Utility Functions. (In: Annals of Statistics, Vol. 4, No. 1, January 1976)

13. Lindley, D. V. Probability and Medical Diagnosis; and The Role of Utility in Decision-Making.

14. Lindley, D. V. Inference for a Bernoulli Process: A Bayesian View; and The Future of Statistics: A Bayesian 21st Century.

15. Shigemasu, Kazuo. Development and Validation of a Simplified M-Group Regression Model. (In: Journal of Educational Statistics, Summer 1976, Vol. 1, No. 2, pp. 157-180)

16. Cantor, Gordon N. Sex and Race Effects in the Community Behavior of Upper-Elementary-School-Aged Children. (ED 115 658)

17. Forsyth, Robert A. Describing What Johnny Can Do. (ED 181 063)

18. Sedere, M. ..., & Feldt, Leonard S. The Sampling Distribution of the Kristof Reliability Coefficient, the Feldt Coefficient, and Guttman's λ_2. (In: Journal of Educational Measurement, Spring 1977, Vol. 14, No. 1, pp. 53-62)

19. Coffman, William E. Those Achievement Tests - What For?

20. Hunt, David E. Teachers are Psychologists, Too: On the Applications of Psychology to Education. (ED 129 651)

21. Johnson, Sylvia T. Self-Concept in a School Setting: Construct Validations by Factor Analysis. (ED 142 573)

22. Coffman, William E., & Shigemasu, Kazuo. Appraising School Effectiveness Using a Bayesian Method. (ED 159 220)

23. Coffman, William E. Classical Test Development Solutions. (ED 175 913)

24. Olsen, Scott A. An Investigation of Possible Test Floor Effects by Comparing Score Distributions for Students Tested with Two Different Levels of a Test Battery at Four Different Times over a Two-Year Interval. (ED 175 914)

25. Haebara, Tomokazu. A Method for Investigating Item Bias Using Birnbaum's Three-Parameter Logistic Model.

26. Coffman, William E., & Olsen, Scott A. The First Two Years of PLAN*-- An Evaluation of Program Impact.

27. Haebara, Tomokazu. Equating Logistic Ability Scales by a Weighted Least Squares Method.

28. Mayekawa, Shin-ichi, & Haebara, Tomokazu. Estimation of the Reliability of a Test Consisting of More Than Three Congeneric Parts.